SortMeRNA User Manual

Evguenia Kopylova
evguenia.kopylova@lifl.fr

January 2013

Contents

1 Introduction **3**

2 Installation **3**

2.1 Required `g++` compiler version 3

2.1.1 Ubuntu Linux OS instructions 3

2.1.2 Mac OS X instructions (using MacPorts) 4

2.2 Install from source code 6

2.3 Install from precompiled code 7

3 Databases **8**

4 How to run SortMeRNA **8**

4.1 Index the rRNA database: command ‘`buildtrie`’ 8

4.1.1 Example 1: buildtrie 9

4.2 Filter reads against the indexed rRNA database: command ‘`sortmerna`’ 10

4.2.1 Example 2: sortmerna on multiple databases 11

4.2.2 Example 3: sortmerna on paired-end reads 12

5 SortMeRNA parameters **14**

1 Introduction

Copyright (C) 2012 Bonsai Bioinformatics Research Group
(LIFL - Université Lille 1), CNRS UMR 8022, INRIA 2012

SortMeRNA is a software designed to filter metatranscriptomic reads data. It takes as input a file of reads (fasta or fastq format) and an rRNA database file, and sorts apart the accepted reads and the rejected reads into two files specified by the user.

For questions & help, please contact:

1. Evguenia Kopylova	evguenia.kopylova@lifl.fr
2. Laurent Noe	laurent.noe@lifl.fr
3. Helene Touzet	helene.touzet@lifl.fr

2 Installation

2.1 Required g++ compiler version

Tested platforms: Ubuntu Linux OS 12.04 and Mac OS X 10.6.8.

SortMeRNA uses features of the `C++11` standard which are available for GCC compiler versions **4.3** and later. To check your compiler version, type '`g++ --version`'. For compiler versions less than 4.3, please follow *Subsection 2.1.1* for Ubuntu Linux OS or *Subsection 2.1.2* for Mac OS X, to configure or install a new compiler.

2.1.1 Ubuntu Linux OS instructions

(1) Check whether you have other `g++` compiler versions installed (4.3 or later), type '`g++`' then press the tab key →,

```
> g++ (followed by the tab key)
```

(2) If the list output in Step (1) does not include a `g++` compiler of version 4.3 or later, install one (SortMeRNA was tested with `g++-4.4`, `g++-4.5` and `g++-4.6`),

```
> sudo apt-get install g++-4.4
```

(3) Determine the full directory path of where `g++-4.4` was installed (in the example below, the path is `/usr/bin/g++-4.4`, but others are also possible under different system configurations),

```
> which g++-4.4
/usr/bin/g++-4.4
```

(4-a) To select `g++-4.4` only for compiling SortMeRNA, the user must specify the full path of the compiler in the Makefile,

```
change the line:
  > CC = g++
to:
  > CC = /usr/bin/g++-4.4 (the full directory path from Step (3) above)
```

(4-b) To set `g++-4.4` as a default compiler, do not edit the Makefile but run the commands,

```
  > which g++                                (find the path of the ‘g++’ command)
    /usr/bin/g++
  > sudo rm /usr/bin/g++                     (remove the symbolic link of ‘g++’)
  > sudo ln -s /usr/bin/g++-4.4 /usr/bin/g++   (create a new symbolic link from ‘g++’
                                              to ‘g++-4.4’)
  > g++ --version                            (should now be 4.4)
```

(5) The user may now run the ‘`make`’ command in Step (2) of Section 2.2.

2.1.2 Mac OS X instructions (using MacPorts)

(1) Check whether you have other `g++` compiler versions installed (4.3 or later), type ‘`g++`’ then press the tab key →,

```
  > g++ (followed by the tab key)
```

(2) If the list given by Step (1) does not include a `g++` compiler of version 4.3 or later, install one (SortMeRNA was tested with `gcc43, gcc44, gcc45, gcc46` and `gcc47`),

```
  > sudo port selfupdate
  > sudo port upgrade outdated
  > sudo port install gcc44
```

(3-a) To select `gcc44` only for compiling SortMeRNA, the user must specify the full path of the compiler in the Makefile,

```
change the line:
  > CC = g++
to:
  > CC = /opt/local/bin/g++-mp-4.4 (the full directory path of the installed compiler)
```

(3-b) To set `gcc44` as a default compiler, do not edit the Makefile but run the commands,

```
  > which g++                               (find the path of the ‘g++’ command)
    /opt/local/bin/g++
  > sudo rm /opt/local/bin/g++              (remove the symbolic link of ‘g++’)
  > sudo ln -s /opt/local/bin/g++-mp-4.4 \  (create a new symbolic link from ‘g++’
               /opt/local/bin/g++            to ‘g++-mp-4.4’)
  > g++ --version                           (should now be 4.4)
```

(4) The user may now run the ‘`make`’ command in Step (2) of Section 2.2. If linking errors occur

such as ‘library not loaded’ or ‘undefined symbols for architecture’, try to upgrade your current compiler version to establish correct links,

```
  > sudo port -n upgrade --force gcc44
```

2.2 Install from source code

1. Download `sortmerna.tar.gz` from `http://bioinfo.lifl.fr/RNA/sortmerna`
2. Extract the source code package into a directory of your choice (must have permissions to read and write) and build the executable files `sortmerna` and `buildtrie`,

   ```
   > tar -zxvf sortmerna.tar.gz
   > cd sortmerna
   > make
   ```

   (**note:** The `g++` compiler version must be 4.3 or later for '`make`' to work. Please see Subsection 2.1 to configure or install a new compiler if the command '`which g++`' returns a version less than 4.3)
3. SortMeRNA indexes a database by writing its contents to the '`/automata`' folder, which is initially located in the directory '`/sortmerna/automata`'. It is essential to set the `$SORTMERNADIR` environmental variable to the path where the '`/automata`' folder is found, so that the program may read and write from it. The user may move the '`/automata`' folder to a workspace with larger memory, separately from the directory '`/sortmerna`'. To find the path of the '`/automata`' folder, go into the directory where it is located and type,

   ```
   > pwd
   /some/path/to/sortmerna
   ```

4. Open the ~/.bashrc (or ~/.profile) file in any editor and add the line,

   ```
   > export SORTMERNADIR="/some/path/to/sortmerna/automata"
   > export PATH="$PATH:/some/path/to/sortmerna"
   ```

5. Run the ~/.bashrc (or ~/.profile) file to add the variable `$SORTMERNADIR` and update the variable `$PATH` in the list of environment variables,

   ```
   > source ~/.bashrc
   or
   > source ~/.profile
   ```

6. Check that `$SORTMERNADIR` has been added,

   ```
   > echo $SORTMERNADIR
   /some/path/to/sortmerna  (if it has not been added,
                             this path will be empty)
   ```

7. Check that `$PATH` has been updated with the additional directory search path,

   ```
   > echo $PATH
   /usr/local/bin:/usr/bin:...:/some/path/to/sortmerna
   ```

8. Check the path of the executables,

```
> which sortmerna buildtrie
/some/path/to/sortmerna/sortmerna
/some/path/to/sortmerna/buildtrie
```

9. To begin using SortMeRNA, type '`buildtrie -h`' or '`sortmerna -h`'.

2.3 Install from precompiled code

1. Download,

   ```
   sortmerna_linux_64.tar.gz  or
   sortmerna_linux_32.tar.gz  or
   sortmerna_mac_64.tar.gz
   ```

   from `http://bioinfo.lifl.fr/RNA/sortmerna`

2. Extract the source code package into a directory of your choice (must have permissions to read and write),

   ```
   > tar -zxvf sortmerna_linux_64.tar.gz
   > cd sortmerna_linux_64
   ```

3. SortMeRNA indexes a database by writing its contents to the '`/automata`' folder, which is initially located in the directory '`/sortmerna_linux_64/automata`'. It is essential to set the `$SORTMERNADIR` environmental variable to the path where the '`/automata`' folder is found, so that the program may read and write from it. The user may move the '`/automata`' folder to a workspace with larger memory, separately from the directory '`/sortmerna_linux_64`'. To find the path of the '`/automata`' folder, go into the directory where it is located and type,

   ```
   > pwd
   /some/path/to/sortmerna_linux_64
   ```

4. Open the ~/.bashrc (or ~/.profile) file in any editor and add the line,

   ```
   > export SORTMERNADIR="/some/path/to/sortmerna_linux_64"
   > export PATH="$PATH:/some/path/to/sortmerna_linux_64"
   ```

5. Run the ~/.bashrc (or ~/.profile) file to add the variable `$SORTMERNADIR` and update the variable `$PATH` in the list of environment variables,

   ```
   > source ~/.bashrc
   or
   > source ~/.profile
   ```

6. Check that `$SORTMERNADIR` has been added,

   ```
   > echo $SORTMERNADIR
   /some/path/to/sortmerna_linux_64   (if it has not been added,
                                       this path will be empty)
   ```

7. Check that the `$PATH` has been updated with the additional directory search path,

   ```
   > echo $PATH
   /usr/local/bin:/usr/bin:...:/some/path/to/sortmerna_linux_64
   ```

8. Check the path of the executables,

   ```
   > which sortmerna buildtrie
   /some/path/to/sortmerna_linux_64/sortmerna
   /some/path/to/sortmerna_linux_64/buildtrie
   ```

9. To begin using SortMeRNA, type '`buildtrie -h`' or '`sortmerna -h`'.

3 Databases

SortMeRNA comes prepackaged with 8 databases,

representative database	**id %**	**average id %**	**# seq**	**origin**	**# seq**	**filtered to remove**
silva-bac-16s-database-id85.fasta	85	91.6	8174	SILVA SSU Ref NR v.111	244077	23s
silva-arc-16s-database-id95.fasta	95	96.7	3845	SILVA SSU Ref NR v.111	10919	23s
silva-euk-18s-database-id95.fasta	95	96.7	4512	SILVA SSU Ref NR v.111	31862	26s,28s,23s
silva-bac-23s-database-id95.fasta	98	99.4	3055	SILVA LSU Ref v.111	19580	16s,26s,28s
silva-arc-23s-database-id95.fasta	98	99.5	164	SILVA LSU Ref v.111	405	16s,26s,28s
silva-euk-28s-database-id95.fasta	98	99.1	4578	SILVA LSU Ref v.111	9321	18s
rfam-5s-database-id98.fasta	98	99.2	59513	RFAM	116760	–
rfam-5.8s-database-id98.fasta	98	98.9	13034	RFAM	225185	–

The tool UCLUST was used to reduce the size of the original databases.

id %: members of the cluster must have identity at least this % id with the representative sequence
average id %: average identity of a cluster member to the representative sequence

Remark: The user must first index the fasta database by using the command `buildtrie` and then filter reads against the database using the command `sortmerna`.

4 How to run SortMeRNA

4.1 Index the rRNA database: command '`buildtrie`'

The executable `buildtrie` indexes an rRNA database.

To see the man page for `buildtrie`,

```
> buildtrie -h
```

```
This program builds a Burst trie on an input rRNA database file in fasta format
and stores the material in binary files under the folder ‘automata’

   ./buildtrie --db [path to rrnas database file name {.fasta}] {OPTIONS}

The list of OPTIONS can be left blank, the default values will be used:

   -L      length of the sliding window (the seed)
           (default: 18)

   -F      search only the forward strand

   -R      search only the reverse-complementary strand
           (default: both strands are searched)

   -h      help
```

There are eight rRNA representative databases provided in the ‘`sortmerna/rRNA_databases`’ folder. All databases were derived from the SILVA SSU and LSU databases (release 111) and the RFAM databases using the tool UCLUST. Additionally, the user can index their own database.

4.1.1 Example 1: buildtrie

```
> buildtrie --db ~/sortmerna/rRNA_databases/silva-bac-16s-database-id85.fasta

  Burst trie(s) built in:      36.7594s
  Writing Burst trie forward to silva-bac-16s-database-id85.bursttrief.dat
  Writing Burst trie reverse to silva-bac-16s-database-id85.bursttrier.dat
  Done.
```

The indexed databases (ex. `silva-bac-16s-database-id85.bursttrief.dat`) will be stored in the directory ‘`/some/path/to/sortmerna/automata`’ (the path stored in variable `$SORTMERNADIR`, which was established in Step 1-4 of Subsection 2.2 or Subsection 2.3) and later retrieved by the command `sortmerna`, explained in the following section.

4.2 Filter reads against the indexed rRNA database: command ‘sortmerna’

The executable sortmerna filters rRNA reads against an indexed rRNA database.

To see the man page for sortmerna,

```
> sortmerna -h

  To run SortMeRNA, type in any order after ‘sortmerna’:

      --I      [illumina reads file name {fasta/fastq}]

      --454    [roche 454 reads file name {fasta/fastq}]

      -n        number of databases to use (must precede --db)

      --db     [rrnas database name(s)]
                One database,
                ex 1. -n 1 --db /path1/database1.fasta

                Multiple databases,
                ex 2. -n 2 --db /path2/database2.fasta /path3/database3.fasta
      {OPTIONS}

  The list of OPTIONS can be left blank, the default values will be used:

      --accept      [accepted reads file name]
      --other       [rejected reads file name]
                    (default: no output files are created)

      --bydbs       output the accepted reads by database
                    (default: concatenated file of reads)

      --log         [overall statistics file name]
                    (default: no statistics file created)

      --paired-in   put both paired-end reads into --accept file
      --paired-out  put both paired-end reads into --other file
                    (default: if one read is accepted and the other is not,
                    separate the reads into --accept and --other files)

      -r            ratio of the number of hits on the read / read length
                    (default Illumina: 0.25, Roche 454: 0.15)

      -F            search only the forward strand
      -R            search only the reverse-complementary strand
                    (default: both strands are searched)

      -a            number of threads to use
```

```
                 (default: 1)

    -v           verbose
                 (default: deactivated)

    -h           help

    --version    version number
```

The command `sortmerna` takes as input a list of rRNA databases (in fasta format) and a set of Illumina or Roche 454 reads (in fasta or fastq format), and filters out the reads matching to at least one of the rRNA databases. The user has an option to output the accepted reads into a single file (default), or into multiple files sorted by the closest matching database (add the flag `--bydbs`). The indexed part of the databases created by `buildtrie` is loaded into `sortmerna` independently.

4.2.1 Example 2: sortmerna on multiple databases

```
> sortmerna -n 2
           --db ~/sortmerna/rRNA_databases/silva-bac-23s-database-id98.fasta
                ~/sortmerna/rRNA_databases/silva-bac-16s-database-id85.fasta
           --454 SRR106861-filtered.fasta
           --log bilan
           -a 3
           -v

 WARNING: option '--accept' has been left blank, no output file for accepted reads ..
 WARNING: option '--other' has been left blank, no output file for rejected reads ..

 -------------------------------------------------------
 Welcome to SortMeRNA!
 Copyright (C) 2012 Bonsai Bioinformatics Research Group
 LIFL, Université Lille 1, CNRS UMR 8022, INRIA 2012
 -------------------------------------------------------

 The size of the reads file <33862846> bytes
 will be executed in 1 partial section(s) of size
 <33862846> bytes

 [Partial section # 1]
 ---------------------
 Time to mmap reads and set up pointers:          0.3525s

 Begin analysis of: ./rRNA_databases/silva-bac-23s-database-id98.fasta

 Time to load the Burst trie:                     1.2637s
 Begin parallel traversal ...
 Time of parallel traversal of automata:          7.9790s
```

```
  Begin analysis of: ./rRNA_databases/silva-bac-16s-database-id85.fasta

  Time to load the Burst trie:                    1.9774s
  Begin parallel traversal ...
  Time of parallel traversal of automata:         10.3811s


  Time to output reads to file:                   0.0027s
  Total number of reads matching database:        95205
```

The option '--log bilan' will create an overall statistics file called bilan.log,

```
> cat bilan.log

reads file:    ~/SRR106861-filtered.fasta

total reads:    105873
non-rRNA:       10668
rRNA:           95205
% rRNA:         89.92%

./rRNA_databases/silva-bac-23s-database-id98.fasta           64.4%
./rRNA_databases/silva-bac-16s-database-id85.fasta           25.53%
```

4.2.2 Example 3: sortmerna on paired-end reads

This example illustrates three cases of output for paired-end reads: default, --paired-in and --paired-out.

We use the set6-database.fasta found under Section 3.2 of the Supplementary file (also available online at bioinfo.lifl.fr/RNA/sortmerna/material.php). As for the reads, 5000 Illumina paired-end reads were simulated using MetaSim on set6-database.fasta. The reads were then arranged to have 2475 pairs of rRNA, and the other 25 pairs where exactly one read is rRNA and the other is not (this is possible if one of the reads covers a low complexity region on the rRNA sequence).

Remark: The statistics in the --log file will always give the true number of reads classified as rRNA.

Case 1: We don't care to keep the paired-end order in the output (default).

```
 > sortmerna -n 1
             --db set6-database.fasta
             --I paired-end-5000-reads.fasta
             --accept rrna
             --other nonrrna
             --log bilan
```

```
            -a 1

> cat bilan.log

 Results:
 total reads:   5000
 non-rRNA:      50
 rRNA:          4950
 % rRNA:        99%

set6-database.fasta       99%

> grep -c '>' rrna.fasta nonrrna.fasta
rrna.fasta: 4950
nonrrna.fasta: 50
```

Case 2: We want to accept all pairs where at least one read that hits (`--paired-in`).

```
> sortmerna -n 1
            --db set6-database.fasta
            --I paired-end-5000-reads.fasta
            --accept rrna
            --other nonrrna
            --log bilan
            --paired-in
            -a 1

> cat bilan.log

 Results:
 total reads:   5000
 non-rRNA:      50
 rRNA:          4950
 % rRNA:        99%

set6-database.fasta       99%

> grep -c '>' rrna.fasta nonrrna.fasta
rrna.fasta: 5000
nonrrna.fasta: 0
```

Case 3: We want to reject all pairs where at least one read that hits (`--paired-out`).

```
> sortmerna -n 1
            --db set6-database.fasta
            --I paired-end-5000-reads.fasta
            --accept rrna
            --other nonrrna
            --log bilan
            --paired-out
            -a 1

> cat bilan.log

 Results:
 total reads:   5000
 non-rRNA:      50
 rRNA:          4950
 % rRNA:        99%

set6-database.fasta      99%

> grep -c '>' rrna.fasta nonrrna.fasta
rrna.fasta: 4900
nonrrna.fasta: 100
```

If the reads file is larger than 1GB, then `sortmerna` internally divides the file into partial sections of 1GB and executes one section at a time. Hence, the user can input a file of 15GB and it will be processed by `sortmerna` without having to physically split the file prior to execution.

5 SortMeRNA parameters

There are two parameters in SortMeRNA which the user can moderate: The length of the sliding window s (the seed), and the threshold ratio r of matching windows to the rRNA database for a read to be accepted. Both of these paramaters and their default values are discussed in detail in *Section 2: Parameter Setting* of the supplementary file. The user may adjust the threshold parameter r with the sortmerna command-line option `-r [new ratio]`. To adjust the length of the sliding window s, the user must provide the option `-L [new length (even integer)]` when indexing an rRNA database using the `buildtrie` executable.